STAT 401 Chapter 1

Zepu Zhang
October 25, 2010

1 Relations between variables

1.1 Functional relation: $y = f(x)$

Data points (x_i, y_i) fall exactly on the function curve.

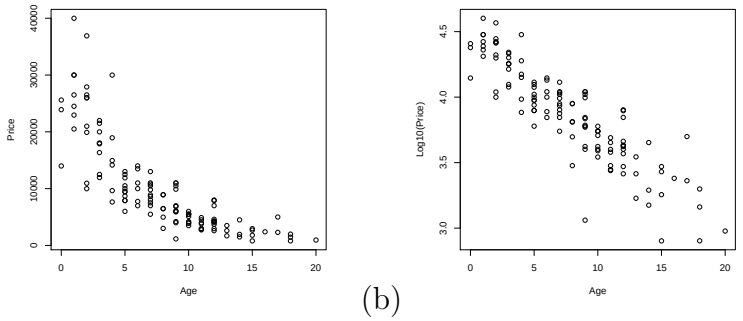
1.2 Statistical relation: $y = f(x) + \epsilon$

For any fixed value x, Y has a random scatter, or fluctuation, about the deterministic function $f(x)$.

Example

1. $Y = \%$ body fat, $X =$ age.
2. $Y =$ cases of beer sold at the liquor store on a given Saturday, $X =$ number of college basketball games being aired that weekend.
3. $Y =$ number of snow shovels sold in a given week, $X =$ average daily temperature for that week.
4. $Y =$ number of miles driven per year, $X =$ urban/rural driver.
5. $Y =$ change in cholesterol level, $X =$ placebo vs. active drug treatment group.
6. $Y =$ direction from beehive at which bee disappears from sight, $X =$ species of bee.
7. $Y =$ actual fish length, $X =$ length as recorded by sonar.

In this course we work with quantitative or qualitative (i.e. categorical) X, and quantitative Y. We do not work with qualitative Y.



(a) (b)

Figure 1: Mazda data. (a) Original. (b) Log transformed.

1.3 Types of studies

Observational

Experimental

Observational study: may deduce correlation, but not causation.

Well-designed, well-run experimental study: may deduce causation.

2 Least squares (LS) estimation

Example data set Mazda Price data: asking prices (Australian dollars) for 124 Mazda cars, along with the age of the car (years). (Reprinted in Data Analysis: An Introduction Based on R, by A. Lee, Auckland: Department of Statistics, University of Auckland.)

Apparently there exists some relation between age and price. The relation is statistical because at a certain age, the price is not uniquely determined.

The relation does not appear to be linear. We log-transform the price, then $\log_{10}(\text{price}) \sim \text{age}$ looks like linear. So we take age as X and $\log_{10}(\text{price})$ as Y.

Suppose a linear relation is a good description for this relationship:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i,$$

where β_0 is intercept and β_1 is slope. They are both regression coefficients. But they are unknown and need to be estimated from the data.

What are the values of β_0 and β_1? In other words, we want to find (or "estimate") a function $y = b_0 + b_1 x$ that is the "best

fit" to this dataset. (Apparently we can't find a "perfect" fit.)

What is the criterion for a "good fit"?

The least squares criterion: the best fit is given by β_0 and β_1 that minimize the following quantity:

$$Q = \sum_{i=1}^{n} (Y_i - \beta_0 - \beta_1 X_i)^2.$$

This is the most commonly used criterion for "best fit", and it makes intuitive sense. In words, we find β_0 and β_1 values that minimize the total squared distances between the fitted values (i.e. $\beta_0 + \beta_1 X_i$) and the observed values (i.e. Y_i).

Denote the solutions by $\hat{\beta}_0$ and $\hat{\beta}_1$. They are found by setting $\frac{\partial Q}{\partial \beta_0} = 0$ and $\frac{\partial Q}{\partial \beta_1} = 0$, that is,

$$\frac{\partial Q}{\partial \beta_0} = \sum 2(Y_i - \beta_0 - \beta_1 X_i)(-1) = 0 \tag{1}$$

$$\frac{\partial Q}{\partial \beta_1} = \sum 2(Y_i - \beta_0 - \beta_1 X_i)(-X_i) = 0 \tag{2}$$

This leads to

$$\begin{aligned} \sum Y_i &= n\hat{\beta}_0 + \hat{\beta}_1 \sum X_i, \\ \sum X_i Y_i &= \hat{\beta}_0 \sum X_i + \hat{\beta}_1 \sum X_i^2. \end{aligned} \tag{3}$$

Its solution is

$$\hat{\beta}_1 = \frac{\sum (X_i - \overline{X})(Y_i - \overline{Y})}{\sum (X_i - \overline{X})^2} = \frac{S_{xy}}{S_{xx}}, \quad \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}. \tag{4}$$

This gives formulas for estimating the regression coefficients using the available data $(x_1, y_1), \ldots, (x_n, y_n)$.

With the Mazda data, we find

$n = 24$,
$\sum X_i = 964$, $\sum Y_i = 479.69$,
$\sum X_i^2 = 10138$, $\sum Y_i^2 = 1872.67$,
$\sum X_i Y_i = 3540.09$.

The LS estimates of the slope and intercept are $b_1 = -0.0715$ and $b_0 = 4.4245$, respectively.

$\hat{\beta}_0$ and $\hat{\beta}_1$ are called "estimators" (i.e. formulas for calculating estimates).

b_0 and b_1 are called "estimates" (i.e. actual values based on actual data).

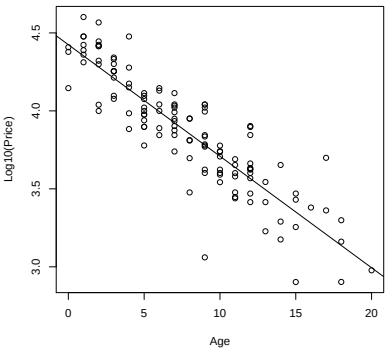


Figure 2: Mazda data: fitted model

An oft-used re-formatting:

$$\begin{aligned}\sum(X_i-\overline{X})(Y_i-\overline{Y}) &= \sum\left(X_iY_i - X_i\overline{Y} - \overline{X}Y_i + \overline{XY}\right)\\ &= \sum X_iY_i - \frac{\sum Y_i}{n}\sum X_i - \frac{\sum X_i}{n}\sum Y_i + n\frac{\sum X_i}{n}\frac{\sum Y_i}{n}\\ &= \sum X_iY_i - \frac{1}{n}\sum X_i\sum Y_i\end{aligned}$$

$$\sum(X_i-\overline{X})^2 = \sum X_iX_i - \frac{1}{n}\sum X_i\sum X_i = \sum X_i^2 - \frac{1}{n}\left(\sum X_i\right)^2$$

Using these relations, the estimator $\hat{\beta}_1$ can be reformatted as

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}} = \frac{n\sum X_iY_i - \sum X_i\sum Y_i}{n\sum X_i^2 - (\sum X_i)^2}$$

3 Simple linear regression models

Consider the following relation between X and Y:

$$Y_i = \beta_0 + \beta_1 X_i + \epsilon_i \tag{5}$$

where

X_i: value of the predictor variable in the ith trial;
Y_i: value of the response variable in the ith trial;
β_0, β_1: regression coefficients (or model parameters), intercept and slope;
ϵ_i: random error in the ith trial.

In regression models the two variables have different roles.
X: predictor, independent variable.
Y: dependent/response variable.

We study the statistical behavior of Y when X assumes a certain value, say x.

This formulation assumes:

1. $E(\epsilon_i) = 0$.
2. $\text{var}(\epsilon_i) = \sigma^2$ is constant across different i.
3. ϵ_i and ϵ_j, where $i \neq j$, are uncorrelated.

The formula (5) along with the preceding assumptions constitute our model for the relation between X and Y.

We call (5) a simple, linear regression model. It's simple because there is only one predictor variable. (This is by definition what we refer to when we talk about "simple linear models"; it has nothing to do with whether such models are indeed simple.) It's linear because the parameters β_0 and β_1 appear in linear forms. (Something like $Y_i = \alpha e^{\beta X_i} + \epsilon_i$ would be a nonlinear model.)

It is much less important that the predictor X happens to appear in linear form in this model.

Further observations on the model:

1. Y_i is the sum of two components: (1) the deterministic term $\beta_0 + \beta_1 X_i$ and (2) the random term ϵ_i. Hence Y_i is a random variable.

2. $$E(Y_i) = E\big(\beta_0+\beta_1 X_i+\epsilon_i\big) = E\big(\beta_0+\beta_1 X_i)+E(\epsilon_i) = \beta_0+\beta_1 X_i$$

   Hence $\beta_0 + \beta_1 X_i$ is the mean of the probability distribution of the random variable $Y_i \,|\, X_i$ (meaning, the random variable Y_i when X is fixed at a certain value X_i).

   $E(Y) = \beta_0 + \beta_1 X$ is called the "regression function", which relates the mean of the probability distribution of Y, for given X, to the level of X.

3. $$\text{var}(Y_i) = \text{var}\big(\beta_0 + \beta_1 X_i + \epsilon_i\big) = \text{var}(\epsilon_i) = \sigma^2$$

   because $\beta_0 + \beta_1 X_i$ is constant. Therefore the variance of Y is constant; it does not change with X.

4. Following the assumed non-correlation between ϵ_i and ϵ_j, we have

   $$\text{cov}(Y_i, Y_j) = \text{cov}(\beta_0+\beta_1 X_i+\epsilon_i,\ \beta_0+\beta_1 X_j+\epsilon_j) = \text{cov}(\epsilon_i,\ \epsilon_j) = 0$$

   hence Y_i and Y_j are uncorrelated.

In sum, model (5) says that Y comes from a distribution with mean $\beta_0 + \beta_1 X$ and variance σ^2, and the Y values corresponding to different X values are uncorrelated.

It is crucial to understand that the model is an assumption about the statistical behavior of the random Y conditional on the value of X. See Figure 16.1(a) on page 680 for an illustration.

The preceding assumptions do not specify the exact probability distribution of $Y \mid X$. (It only states it's a distribution with mean $\beta_0 + \beta_1 X$ and that variance σ^2.) The most common extra step is to assume the distribution is normal, that is

$$Y \mid X \overset{\text{iid}}{\sim} N(\beta_0 + \beta_1 X,\ \sigma^2)$$

With this normal assumption, a lot more things can be done. See maximum likelihood estimation.

4 Least squares estimation (continued)

In this section we do not assume normality for the error terms ϵ_i.

4.1 Estimation of the regression coefficients β_0, β_1

Why are the LS estimators "good"?

—Intuitively the criterion makes sense.
—More importantly, they are unbiased, minimum variance linear estimators.

1. Both estimators $\hat{\beta}_0$ and $\hat{\beta}_1$ are linear functions of the observed Y's. (The X's are regarded as constants, so are not a concern here.)

Exercise Verify this statement.

2. Both are unbiased, meaning

   $$E(\hat{\beta}_0) = \beta_0, \quad E(\hat{\beta}_1) = \beta_1.$$

The expectations are taken wrt repeated sampling of Y's, keeping the set of X's unchanged.

β_0 and β_1 are statistics of the data $\{(X_i, Y_i)\}$, i.e. functions of the data. They are random, changing as we re-sample Y's at the fixed set of X values. As we do this re-sampling, $\hat{\beta}_0$ and $\hat{\beta}_1$ each has a sampling distribution (with certain properties like mean, variance, distribution type). Now we know the distributions are "centered" at their true values.

Exercise Verify this statement.

3. Among all linear, unbiased estimators of β_0 and β_1 (assuming there are alternative estimators), $\hat{\beta}_0$ and $\hat{\beta}_1$ have the smallest sampling variances.

   "Minimum variance" is certainly a nice property. It suggests the fluctuation (due to randomness of the data) of the estimator (around its mean, here the true value, due to unbiasedness) is small, therefore the estimate is more "precise".

   A proof of this statement will be given in Chapter 2; but you are not required to know the proof.

4.2 Interpretation of the regression coefficients

Slope β_1: change of $E(Y)$ for one unit increase of X.

Intercept β_0: value of $E(Y)$ at $X = 0$.

The interpretation of β_1 is more important than β_0:

(1) The slope indicates the association between X and Y (Positive? Negative? How fast does Y tend to change with X?) The mean Y value at the specific X value of 0 is of less interest.

(2) Never extrapolate far out of the data range—If 0 is far out of the range of X values in the data, the linear relation may not apply there, thus we do not know $E(Y)$ at $X = 0$. In that case, β_0 is merely a coefficient in the regression function, which is usable only in the data range.

4.3 Fitted values and residuals

The regression function

$$E(Y) = \beta_0 + \beta_1 X$$

is estimated by

$$\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$$

$\hat{Y}$ is the estimated mean response at a particular X value. (It's "estimated" because the model parameters β_0 and β_1 are estimated; their true values are unknown.)

Corresponding to a predictor value X_i in the data, the value $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i$ is called the fitted value.

The discrepancy between the observed value and the fitted value,

$$e_i = Y_i - \hat{Y}_i$$

is called the residual. Note, if $\hat{Y}_i$ were the true $E(Y_i) = \beta_0 + \beta_1 X_i$, then e_i would be identical to ϵ_i in the model. Now that $\hat{Y}_i$ is only an "estimate" of $E(Y_i)$, the residual e_i is not ϵ_i; rather, it is an estimate of ϵ_i.

The LS estimate minimizes the total squared residuals, that is, $Q = \sum e_i^2$.

Exercise Verify this statement by looking at the definition of Q before.

Corresponding to a predictor value X that is not in the data, the value $\hat{Y} = \hat{\beta}_0 + \hat{\beta}_1 X$ is (sometimes) called the predicted value. The predicted value is actually the estimated mean response at X, i.e. the mean of the distribution of $Y \mid X$. This apparently is the best single-value prediction one can give.

Remember, $(X, \hat{Y})$ is a point on the regression line.

Example Mazda Price data.

4.4 Estimation of the error variance σ^2

An estimate of $\sigma^2 = \text{var}(\epsilon)$ is obtained based on the e_i's. Let the "sum of squared error" be

$$\text{SSE} = \sum e_i^2,$$

then the "mean squared error" is

$$\text{MSE} = \frac{\text{SSE}}{n-2}, \quad \text{denoted by } s^2.$$

The sum is divided by $n-2$ "degrees of freedom". (The quantity SSE has $n-2$ degrees of freedom because it is calculated using the n free-standing Y values and two parameters, b_0 and b_1, which are estimated using the Y's.)

s^2 is an estimator of σ^2. In fact it is an unbiased estimator.

We use $s = \sqrt{s^2}$ as an estimator of σ. (This estimator is not unbiased, but it's not a big problem. We view σ^2 as a major quantity and σ something secondary.)

Note $\sum e_i = 0$, hence $\bar{e} = 0$, then $s^2 = (n-2)^{-1} \sum (e_i - \bar{e})^2$.

Example Mazda Price data. $s^2 = 0.029$.

Exercise Verify that $\sum e_i = 0$.

5 Maximum likelihood estimation (MLE)

In this section we assume normality for the random error terms, that is, $\epsilon \overset{\text{iid}}{\sim} N(0, \sigma^2)$.

Note that the LS estimation does not need this assumption. With this assumption, we have an alternative estimation method—the maximum likelihood estimation (MLE). MLE is very general and fundamental. It is in a sense more formal than LS. (LS needs to make a somewhat subjective choice about what to minimize. In MLE there is no room for such choice.)

With the distribution assumption, we have

$$Y_i \sim N(\beta_0 + \beta_1 X_i, \sigma^2)$$

therefore the probability density of the observed value y_i is

$$p(y_i) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{ -\frac{1}{2} \frac{(y_i - \beta_0 - \beta_1 x_i)^2}{\sigma^2} \right\}.$$

Because the Y_i's are independent of each other, the joint density of the observed data vector $y_i, \ldots, y_n$ is the product of the individual densities:

$$p(y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i)$$

(Remember the X's are kept fixed.) This function contains three unknown parameters: β_0, β_1, σ^2.

The maximum likelihood principle says the parameters should maximize the density of the observed data. (The principle believes such parameter values are the most possible values of the unknown parameters, given the data.)

The principle is so named because the density of the observed data vector, $p(y_1, \ldots, y_n)$, when viewed as a function of the

unknown parameters (now the data are simply known constants), is called the "likelihood" of the parameters, and is denoted by L:

$$L(\beta_0, \beta_1, \sigma^2; y_1, \ldots, y_n) = \prod_{i=1}^{n} p(y_i)$$

To understand the concept of likelihood—Because we have assumed the distribution of Y, we can write out the probability density at particular Y values, in this case, the observed vector-value $(y_1, \ldots, y_n)$. Now this density function contains x's, y's, and β_0, β_1. However, the data are already observed—they are simply constants; they won't change. What are changeable are the unknown parameters β_0 and β_1. Therefore the probability density is a function of β_0 and β_1. If we try different values for β_0 and β_1, we'll get different density values. The ML principle says the β_0 and β_1 values that make the density as large as it can be are the best estimates for the parameters.

We find the MLE by maximizing $\log L$. (Maximizing $\log L$ is usually easier than maximizing L because the former is a sum whereas the latter is a product.) Setting

$$\frac{\partial \log L}{\partial \beta_0} = 0, \ \frac{\partial \log L}{\partial \beta_1} = 0, \ \frac{\partial \log L}{\partial \sigma^2} = 0,$$

the solution is

$$\hat{\beta}_1 = \frac{S_{xy}}{S_{xx}}, \ \hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X}, \ \hat{\sigma^2} = \frac{\sum e_i^2}{n}.$$

Exercise Derive the ML estimates for β_0, β_1, and σ^2.

Comparison of LS and ML estimates

	LS	ML
Estimate for β_0, β_1	unbiased	(same;) unbiased
Estimate for σ^2	unbiased	biased (underestimates σ^2)
Restrictions on errors ϵ_i	uncorrelated	independent
Distribution of errors ϵ_i	(mean 0, finite variance)	$N(0, \sigma^2)$

- Least squares and maximum likelihood give identical estimates for β_0 and β_1, so it makes no difference which method we say we are using.

- We use the unbiased estimates for σ^2, that is, the MSE or s^2. (This is the customary choice. When n is large the difference is small.)

- From now on, we require the assumption that $\epsilon_i \overset{\text{iid}}{\sim} N(0, \sigma^2)$. (If we only assume the weaker conditions from Least Squares, i.e. uncorrelated errors with mean 0 and variance σ^2, we can not go much further than what we have done so far.)

- Note that 'uncorrelated' and 'independent' are both statements about the (lack of) relationship between observations. In general, being 'independent' is a stronger (more restrictive) condition than being 'uncorrelated.' For normally distributed random variables, 'uncorrelated' and 'independent' are equivalent conditions.

Remark on terminology:
"estimation" refers to the procedure, topic, technicque, etc.
"estimator" refers to the formula, which is a statistic (i.e. function) of data.
"estimate" is the actual value obtained after you plug in the actual data.

6 Steps in regression analysis

1. Exploratory analysis: the data, the subject matter,...

2. Decide on a regression model: compare and choose from multiple feasible models, select predictor variables to use,...

3. Estimate model parameters: point estimates and quantification of uncertainties...

4. Use the estimated model: predict un-observed responses,...